Building a gold standard to construct search filters: a case study with biomarkers for oral cancer*† EC

John J. Frazier, DMD, MSPH; Corey D. Stein, MS; Eugene Tseytlin, MS; Tanja Bekhuis, PhD, MS, MLIS, AHIP

See end of article for authors' affiliations.

DOI: http://dx.doi.org/10.3163/1536-5050.103.1.005

Objective: To support clinical researchers, librarians and informationists may need search filters for particular tasks. Development of filters typically depends on a ''gold standard'' dataset. This paper describes generalizable methods for creating a gold standard to support future filter development and evaluation using oral squamous cell carcinoma (OSCC) as a case study. OSCC is the most common malignancy affecting the oral cavity. Investigation of biomarkers with potential prognostic utility is an active area of research in OSCC. The methods discussed here should be useful for designing quality search filters in similar domains.

Methods: The authors searched MEDLINE for prognostic studies of OSCC, developed annotation guidelines for screeners, ran three calibration trials before annotating the remaining body of citations, and measured inter-annotator agreement (IAA).

Results: We retrieved 1,818 citations. After calibration, we screened the remaining citations (n=1,767; 97.2%); IAA was substantial (kappa=0.76). The dataset has 497 (27.3%) citations representing OSCC studies of potential prognostic biomarkers.

Conclusions: The gold standard dataset is likely to be high quality and useful for future development and evaluation of filters for OSCC studies of potential prognostic biomarkers.

Implications: The methodology we used is generalizable to other domains requiring a reference standard to evaluate the performance of search filters. A gold standard is essential because the labels regarding relevance enable computation of diagnostic metrics, such as sensitivity and specificity. Librarians and informationists with data analysis skills could contribute to developing gold standard datasets and subsequent filters tuned for their patrons' domains of interest.

INTRODUCTION

The biomedical literature is ever growing and, therefore, poses a serious challenge to researchers and clinicians who need to find relevant literature. For example, MEDLINE [1], the US National Library of Medicine's (NLM's) premier bibliographic database, includes more than 21 million references, to which 2,000 to 4,000 references are added 5 times a week [2]. Additionally, more than 2 million biomedical articles were published in North America from 2000–2009, a 42% increase when compared to the previous decade [3]. Clearly, finding what one needs in a large database can be a daunting task, especially for the naïve user [4]. The naïve user will generally enter keywords to retrieve a list of citations, many of which may be relevant. Unfortunately, the length of the list could be in the thousands. At this point, the user might try other combinations of keywords to increase precision. However, suboptimal search strategies are a major deterrent to satisfying the user's information needs [4].

A more effective strategy for searching electronic databases is to use search filters or queries, also known as hedges [5]. Librarians or informationists typically develop filters to target a specific domain, given criteria for inclusion and exclusion. The evolution of methods for developing filters may be characterized by generation. First-generation filters are developed by librarians based on their expertise in searching, but with no empirical validation; second-generation filters are similarly developed, except that validation is based on a gold standard or reference dataset; and third-generation filters are based on statistical approaches and automated methods beyond computing performance measures against a gold standard [6]. The methods for each generation are not mutually exclusive. For both second- and third-generation filters, development typically involves computing diagnostic metrics, such as sensitivity, specificity, and positive predictive value. To do so requires a gold standard dataset with known relevant or positive cases. The accuracy of the labels regarding relevancy is reasonably high, given human judgments, and therefore, the gold standard can serve as a reference against which performance of filters can be assessed [7].

One way to develop a gold standard is to manually screen a broad-based group of citations, some of which meet the user's information needs [8]. Different approaches exist for collecting the initial set. For example, in reviewing twenty papers on development

* The US National Library of Medicine of the National Institutes of Health (NIH) partially supported this research (grant numbers 5T15-LM007059 and 5R00LM010943). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

† Based on a presentation delivered at the annual meeting of the American Academy of Oral & Maxillofacial Pathology; St. Augustine, FL; April 25–30, 2014.

EC A supplemental appendix is available with the online version of this journal.

of methodological filters, Jenkins et al. state that the most common way to collect citations for a gold standard is by hand searching alone or in combination with database searching [6]. Hand searching involves one or more persons manually reviewing, to specified criteria, each article in a set of selected journals. Database searching involves using filters to retrieve citations that meet specified criteria from electronic databases. However, other methods exist, such as collecting references from relevant systematic reviews [7]. Given a retrieval set, developing the gold standard typically involves manual screening by at least two annotators who independently judge the relevance of each citation (including title and abstract) [9; 10, p. 28; 11, p. 413]. Substantial annotator agreement is critical. When agreement is high, the annotated corpus (collection of citations) can serve as a gold standard. If agreement is low, the developers must analyze the reasons for discrepancies and then try to improve agreement, often by modifying guidelines for annotators. After screening, all citations will have been labeled (annotated) with respect to relevancy and together can serve as the gold standard. In this study, the gold standard includes citations labeled as relevant because they are about studies of potential prognostic biomarkers for oral squamous cell carcinoma (OSCC).

OSCC is the most common form of cancer affecting the oral cavity, representing approximately 94% of all malignancies in the mouth [12]. Overall mortality has not changed much in the past 30 to 40 years. Just 50%–60% of newly diagnosed patients are alive in 5 years [12]. With advances in molecular biology and personalized medicine [13, 14], biomarkers of other forms of cancer have been shown to be useful in treatment and prognosis [13, 15]. However, not one marker has proved to be prognostic or diagnostic for OSCC, even though thousands of novel markers have been studied [16]. In the case of OSCC, biomarkers are greatly needed to help improve prognosis and treatment modalities.

An important aspect of discovery and therapeutic use of biomarkers is that investigators and clinicians must be aware of current research, including systematic reviews. However, Choong and Tsafnat report that just 3% of 147,000 systematic reviews indexed in PubMed between 1990 and 2011 are dedicated to biomarkers, even though the growth rate for biomarker publications exceeds that of non-biomarker publications [17]. Thus, biomarker filters tuned to the domain of interest are essential for both discovery and evidence synthesis.

To check that validated biomarker filters do not already exist for our purposes, the authors searched several resources (see below). We were unable to find a validated prognostic biomarker filter that we could satisfactorily combine with terms for OSCC.

This paper presents generalizable methods for developing a gold standard dataset. More specifically, we describe a broad search strategy to retrieve citations from MEDLINE about OSCC and prognosis, selection and use of an annotation tool to label biomarker citations, development of annotator guidelines based on inclusion and exclusion criteria, and the calibration and consensus process. A broad search ensures high recall (i.e., is sensitive). An annotation tool is software with an interface to semi-automate the screening task, as well as record human relevancy judgments in an underlying database, such as MySQL. Annotator guidelines standardize the screening task by defining rules for determining relevance. Calibration refers to a period prior to annotation of the dataset when small trials inform revision of the guidelines to improve consensus (inter-annotator agreement). Figure 1 displays a flowchart depicting the steps involved in constructing a gold standard.

METHODS

Literature review for reports of biomarker filter studies

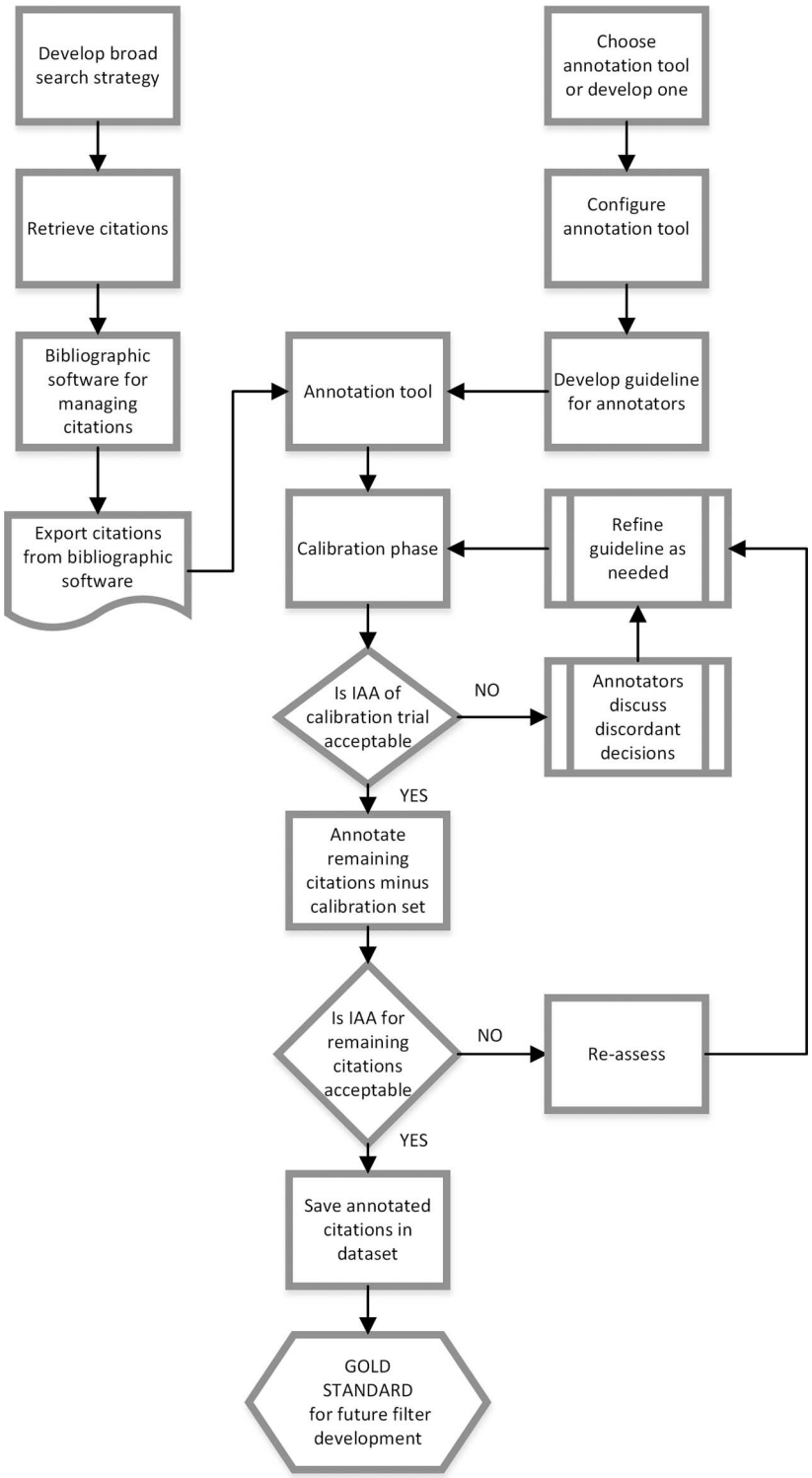
To look for reports in the research literature, we searched several resources for studies on development of biomarker filters, including MEDLINE; Embase [18]; Library, Information Science and Technology Abstracts (LISTA) [19]; and the InterTASC Information Specialists' Sub-Group's (ISSG's) Search Filter Resource [20]. We consulted a librarian at the University of Pittsburgh Health Sciences Library System, who wrote a MEDLINE query that we translated for Embase and LISTA. The first author (Frazier) screened all citation titles: If a title seemed relevant, he read the abstract. If the abstract appeared relevant, he read the full-text report to determine if it pertained to biomarker filter development. Any questions regarding relevancy were discussed with the senior author (Bekhuis).

Search strategy

To develop a gold standard dataset, we first developed a MEDLINE search strategy *without* terms for biomarkers to retrieve citations for prognostic OSCC studies. Note that citations include bibliographic information, such as title, author, publication source, and abstract, as well as Medical Subject Headings (MeSH) added by professional indexers [21]. We reasoned that this retrieval set would include a subset of relevant studies necessary for future development and validation of search filters to find prognostic biomarker studies in this domain.

Both the total number of citations and percentage of relevant citations needed to create a gold standard dataset for this domain are unknown, although we did find some guidance. For example, Pustejovsky and Stubbs suggested reviewing the size of retrieval sets in similar studies [10]. In our case, two recent studies reported retrieval sets of 2,255 and 1,204 citations for developing search filters [22, 23]. We used these numbers as a threshold for this study, namely, our retrieval set had to include more than 1,200 citations. Interestingly, in an analytical study of sample sizes required for filter development, Yao et al.

Figure 1
Flowchart for development of a gold standard dataset



IAA=inter-annotator agreement.

reported that to develop or update a filter for prognostic studies with good sensitivity, one should have between 1,534 and 1,065 methodologically sound prognostic articles (95% confidence intervals with widths equal to 0.05 or 0.06, respectively) [24]. The threshold that we selected falls in this range. We did not set a threshold for the percentage of relevant citations.

Table 1
MEDLINE search to retrieve prognostic studies of oral squamous cell carcinoma in humans

Step	Search string	No of hits
#1	prognosis/broad[filter] AND ((“mouth”[MeSH Terms] OR “mouth” [All Fields] OR “oral”[All Fields]) AND (“carcinoma, squamous cell”[MeSH Terms] OR (“carcinoma”[All Fields] AND “squamous”[All Fields] AND “cell”[All Fields]) OR “squamous cell carcinoma”[All Fields]))	6,225
#2	#1 NOT medline[sb]	308
#3	#1 NOT #2	5,917
#4	#3 AND (“2008/05/16”[PDat] : “2013/05/16”[PDat])	1,842
#5	#4 AND “humans”[MeSH Terms]	1,818

To determine eligibility for this investigation, we used the following criteria: citations had to (i) represent studies about primary squamous cell carcinoma of the oral cavity in humans and prognosis; (ii) have a PubMed ID (PMID) and MeSH terms; and (iii) have been published between May 16, 2008 and May 16, 2013. The PMIDs serve as unique identifiers for citations, which is essential for future filter development and evaluation. Additionally, the MeSH terms are invaluable for future development, which will entail text mining. The date limits ensured that citations were current. Moreover, because the growth rate of biomarker publications is exponential [17], restricting the retrieval set to the most recent five years ensured feasibility of this study.

To retrieve high-quality prognostic studies, we included the prognostic filter developed by the Haynes group as part of our search strategy [4]. This filter is available in PubMed as a clinical query; it appears in our search as *prognosis/broad*, where *broad* indicates that the prognosis filter was tuned for optimal sensitivity (Table 1). To find studies of OSCC, we used MeSH terms and free text related to the oral cavity and squamous cell carcinoma. Recall that entry terms are synonyms, alternate forms, or closely related terms; they map to preferred MeSH terms, thereby enhancing retrieval. For example, ‘‘Oral Cavity’’ is an entry term that maps to ‘‘Mouth,’’ the preferred MeSH term. Interestingly, PubMed expands a query for ‘‘Oral’’ by adding the MeSH term for ‘‘Mouth,’’ even though ‘‘Oral’’ is not an entry term. Thus, the query ‘‘Oral’’ returns citations indexed with ‘‘Mouth.’’ Free text refers to words or phrases appearing in the citation. Free text combined with MeSH terms in a query can retrieve additional citations of interest. For example, we used a MeSH term and free text for *mouth* and *oral* to retrieve citations for the oral cavity. Thus, by using Boolean operators to combine terms for the oral cavity, ‘‘Squamous Cell Carcinoma’’ and ‘‘Prognosis,’’ we retrieved records for prognostic studies of OSCC likely to include records for studies of interest, in other words, on potential prognostic biomarkers. Limits for fully processed articles (indicated by medline[sb]), publication dates, and humans corresponded to the remaining inclusion criteria (Table 1).

We managed the retrieval set in Endnote® v.6 [25]. From Endnote, we exported citations in MEDLINE format as a single text file, which we used as input for an in-house annotation tool.

Annotation tool

In 2010, researchers in the Department of Biomedical Informatics (DBMI), University of Pittsburgh School of Medicine, developed the Gastrointestinal Annotation Tool (GIANT) for records about gastrointestinal conditions. This web-based tool enables manual annotation of textual reports in the medical archival retrieval system (MARS) for research based on University of Pittsburgh Medical Center (UPMC) records. Recently, GIANT has proved useful for annotating textual records for other research endeavors. GIANT is not an open source tool; however, permission to use GIANT may be granted for projects outside of UPMC.

GIANT allows the user to import reports and simultaneously display side-by-side a set of questions for the annotators. GIANT stores the annotations and screening decisions in a MySQL database for subsequent analysis. For this investigation, a DBMI systems programmer processed each citation’s title, abstract, and indexing terms as if the components were part of a MARS clinical report, in effect, repurposing GIANT for our task. The GIANT interface is shown in Figure 2.

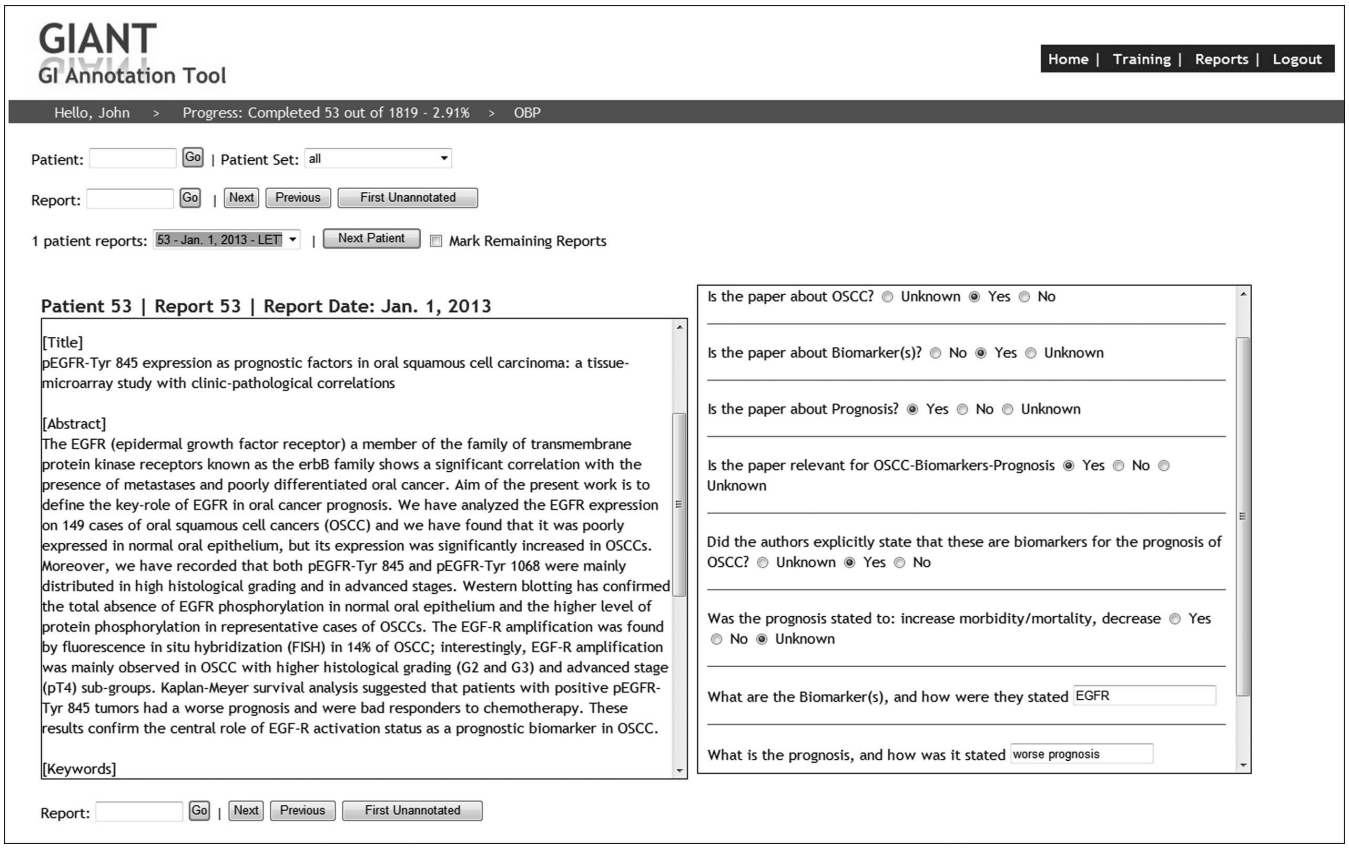
Two annotators, an oral pathologist (Frazier) and a dental informatician (Stein), responded to questions after reading each citation. In designing the questions, we had three major goals: (i) to establish whether titles and abstracts contained information regarding OSCC, biomarkers, or prognosis, considered separately; (ii) to determine whether citations appeared to be about potential prognostic biomarkers for OSCC; and (iii) to extract textual information useful for future filter development involving automated recognition of named biomarkers or analysis of prognostic language.

Annotation guidelines

Annotators considered both titles and abstracts; in a few cases, they considered just titles when the abstract was missing. To be tagged as a relevant citation, responses to all of the following four questions had to be *yes:* Is the paper about OSCC? Is the paper about biomarkers? Is the paper about prognosis? Is the paper relevant for OSCC and biomarkers and prognosis?

If the response to any of the four questions was *no*, the citation was tagged as not relevant. A more detailed description of the annotation guidelines appears in the appendix (online only).

Figure 2
Interface for the Gastrointestinal Annotation Tool, now referred to as the General Information Annotation Tool



For this task, a "patient report" is a portion of the complete citation for a scientific article; keywords are National Library of Medicine Medical Subject Headings assigned by PubMed indexers (not displayed).

Calibration process

One requirement for developing a gold standard is to reduce bias by enlisting the help of two or more annotators who then make decisions individually as to the relevance of a record, depending on the research question. To instill confidence that the articles selected and marked as relevant are accurately labeled, one would expect good agreement among the annotators. However, agreement is typically poor at the outset because annotators are familiarizing themselves with the annotation tool, procedures, definitions, and so on, and therefore need an adjustment period. This period is known as calibration.

One measure of agreement between annotators or inter-annotator-agreement (IAA) is Cohen's kappa statistic. Pusterjovsky and Stubbs provide a rationale for computing IAA and the use of Cohen's kappa statistic [10]. Landis and Koch devised a set of guidelines to interpret agreement metrics such as Cohen's kappa [26].

The calibration procedure was as follows. Two annotators used GIANT to independently annotate blocks of citations. After each block, the annotators discussed their judgments regarding the citations. Discussion included talking about why they agreed on the questions and detailed discussions about why they did not. The annotators then came to consensus regarding disagreements. Following discussion, we adjusted procedures, definitions, and methods to increase the number of concordant judgments.

We calculated Cohen's kappa after each block to assess IAA. The first block (n=11) was screened primarily to plan logistics and to familiarize the annotators with the process. Additionally, the choice of 11 records ensured inclusion of a citation with no abstract and subsequent discussion as to how to standardize their consideration of such citations. The annotators then screened 2 additional blocks of 20 citations each, identified records where judgments were discordant, and reached consensus by discussion. We conducted calibration runs until the annotators reached good agreement. Good agreement was defined as kappa≥0.61, using Landis and Koch's guideline. Additionally, we looked for an improving trend for kappa over runs and apparent stability as stopping criteria.

RESULTS

Literature review for reports of biomarker filters

We retrieved 222 citations from 4 resources: MEDLINE (167, 75.2%), Embase (54, 24.3%), LISTA (1, 0.5%), and ISSG (0). Thirty-six titles (MEDLINE, 25;

Table 2
Confusion matrices and kappa values for calibration trials

		Annotator 1					
		Trial 1		Trial 2		Trial 3	
		Yes	No	Yes	No	Yes	No
Annotator 2	Yes	0	2	4	0	3	0
	No	0	9	1	15	0	17
	Kappa		0.00		0.86		1.00

Table 3
Confusion matrix and kappa value for corpus (minus citations used for calibration)

		Annotator 1	
		Corpus	
		Yes	No
Annotator 2	Yes	430	122
	No	53	1,162
	Kappa	0.76	

Embase, 11) were relevant for reading the abstracts. After de-duplication of abstracts (n=1), 6 papers were read (MEDLINE, 4; Embase, 2). No reports of biomarker filters were identified.

Search strategy

We retrieved 1,818 citations on May 16, 2013, from MEDLINE via PubMed.

Calibration

The calibration process involved annotating 51 citations. Three trials with blocks of 11, 20, and 20 citations produced kappa values of 0.00, 0.86, and 1.00, respectively. IAA was deemed good, as kappa exceeded 0.61 on the second and third trials, and showed a pattern of increasing agreement and stability. Table 2 presents the confusion matrices and kappa scores for the 3 trials. The confusion matrices reflect annotator judgments as to relevance regarding prognostication of OSCC using biomarkers.

Annotation of the corpus

After calibration, the annotators screened the remaining 1,767 (97.2%) citations. Table 3 displays the confusion matrix for the corpus minus the citations used for calibration. The annotators disagreed on 175 of 1,767 (9.9%) citations. Kappa=0.76, which is substantial according to Landis and Koch [26]. The gold standard dataset includes 497 (27.3% of 1,818) relevant citations for studies of potential prognostic biomarkers in OSCC.

DISCUSSION

As a first step in the development of a gold standard dataset, citations that broadly represent scientific articles of interest must be retrieved. Then, the task is to screen the dataset, searching for citations pointing to relevant articles. The resulting collection is about a broad topic that includes a subset of citations representing studies of particular interest. In our case, the retrieval set is about prognostic studies of OSCC with a subset of citations for studies about potential prognostic biomarkers. The latter are the positive or relevant citations that we wish to retrieve when developing and evaluating future filters.

For this study, we restricted our searches to MEDLINE. To identify prognostic studies of OSCC, we integrated the filter for finding scientifically sound prognostic studies developed by the Haynes group into our filter [4]. Although the Haynes filter for prognosis does not perform as well as other MEDLINE filters, such as for therapy or diagnosis, it is currently the best available [27]. Other research teams are actively engaged in developing better prognostic filters [23].

We wished to lay the groundwork for developing a good prognostic biomarker filter for studies of OSCC [28, 29]. We therefore added MeSH terms and free text to the Haynes filter. Oddly, no MeSH term exists for ''Oral'' or for ''Oral Squamous Cell Carcinoma,'' the condition we wish to study. Therefore, to find relevant citations, we used Boolean operators to combine the MeSH terms for ''Mouth'' and ''Carcinoma, Squamous Cell'' with free text terms for oral. Note that an entry term for ''Mouth'' is ''Oral Cavity'' and therefore is appropriate.

To record screening decisions and simultaneously develop a MySQL structured dataset, we elected to use an in-house annotation tool that is web based and user friendly. In this way, we fostered collaboration even though the annotators worked in different cities. However, GIANT is not appropriate if users want to annotate spans of text within a citation as there is no mechanism for tagging text directly. GIANT does enable answering questions about the content of a citation by using radio buttons, check boxes, or text boxes. In our case, two questions have fields for text extracted from titles or abstracts—mainly for future research involving natural language processing. The only way to capture this information in GIANT is to paste the text into the text fields. Nevertheless, the annotators did not find this to be particularly burdensome.

We did have some concern that the presence of indexing terms in the display of a record could bias annotator judgments because their task was to respond to questions solely based on information in the titles and abstracts. We imposed this restriction because indexing is not consistent across indexers [30–32]. For example, knowing that a citation had been indexed with the MeSH term, ''Biological Markers,'' could have affected their judgments in a way that was incompatible with the guidelines. After some discussion, masking the terms in GIANT was deemed unnecessary as the annotators reported that it was simple to ignore them, presumably because neither was trained in library science.

The calibration process required just three trials before acceptable and stable agreement between the

annotators was achieved. The lack of agreement in the first trial can be attributed to the developmental nature of the first version of the annotator guidelines and the inexperience of one of the annotators regarding OSCC. We therefore modified the guidelines for better understanding after the first and second trials, which led to improved consensus.

Disagreement about OSCC was rare and, when it occurred, had to do with the use of ''carcinoma'' as a synonym for ''squamous cell carcinoma.'' ''Carcinoma'' is a generic term for any epithelial malignancy, while ''squamous cell carcinoma'' is a distinct entity. Because an extremely high percentage of carcinomas in the oral cavity are of the squamous cell type, the use of carcinoma to denote squamous cell carcinoma is common. Excluding citations where scientists referred only to carcinoma would lead to a loss of potentially relevant papers. We, therefore, included them when the annotators felt reasonably sure from the context that the authors were referring to squamous cell carcinoma. Additionally, disagreements occurred when the non-pathologist annotator included citations mentioning premalignant lesions. However, the inclusion criteria stipulated that studies had to be about patients who have the disease.

Another source of disagreement had to do with the definition of prognosis. In general, most of the articles concerning prognosis were straightforward in their indication of the disease and/or patients' progression. Phrases with modifiers such as *poorer prognosis, worse outcome,* and *decreased survival* caused few problems. Discrepancies arose when prognosis was implied. For example, the following scenarios were easy for the oral pathologist to interpret, but challenging for the annotator who was not a pathologist:

i. When authors referred to known prognostic indicators such as *tumor metastasis* and *tumor grade* or wrote ''*there is a statistically significant difference in metastasis,*'' the pathologist inferred that, in general, the prognosis was poorer.

ii. When authors wrote about the biological behavior of squamous cell carcinoma but did not directly relate this to the patient, the pathologist inferred that this biological behavior would lead to differences in prognosis. For instance, the authors might state that a biomarker increased the growth rate of malignant squamous cells.

If we were to exclude citations that fell into one of the two scenarios above, we would miss articles of interest concerning potential prognostic biomarkers of OSCC. We, therefore, included them if, after discussion, both annotators agreed.

Another source of disagreement was due to one annotator consistently assigning a *Yes* for prognosis regarding a surgical procedure. However, these articles were about the outcome of reconstructive surgical procedures in patients with OSCC and not about prognosis of the disease.

In sum, once types of disagreements were identified, rapid resolution followed during subsequent discussions to reach consensus.

Strengths of this study

Two annotators screened the citations, which reduces bias and human error. One of the annotators is a board-certified, oral pathologist (Frazier), and both received formal training in an NLM-funded informatics program in the Department of Biomedical Informatics, University of Pittsburgh School of Medicine. This combination brought domain and technical expertise to the project. The pathologist's participation was invaluable because he formatively revised the guidelines, given the difficulties in interpretation that the second annotator experienced during the calibration trials. He also clarified technical points when disagreements arose in screening the corpus. Substantial annotator agreement suggests that both the guidelines and selection criteria were well developed.

The measure of IAA (overall kappa=0.76) is considered substantial on the Landis and Koch scale. The implication is that the OSCC gold standard dataset is likely to be of high quality and will be a useful reference standard for subsequent filter development.

Limitations

This study restricted identification of OSCC studies and potential prognostic biomarkers to information in titles and abstracts in MEDLINE citations. By not screening full-text articles, information appearing in the body of a paper, but not in a citation, would have been missed. Note that while search filters are typically used to retrieve relevant *citations,* we are concerned here with the accurate identification of positive cases in the retrieval set. Thus, information external to the set of citations could have been useful, especially when citations have no abstracts. In our retrieval set, 45 (2.5%) citations were ''empty,'' inasmuch as they had titles but were missing abstracts. It is very unlikely that a title contains enough information to determine relevancy. Even though the percentage of empty citations was small, it is possible that relevant studies are somewhat underrepresented in our gold standard. Nevertheless, the OSCC gold standard has relatively more citations with abstracts (97.5%) compared to MEDLINE (84% since 2010) [2].

Another limitation is that we deliberately kept the number of selection criteria to a minimum. We did this to focus the attention of the annotators on the main criteria for relevancy, but we might have inadvertently omitted important ones. Finally, we limited our searches to MEDLINE, in part because it is an important repository of biomedical citations, but also because it is freely available.

CONCLUSIONS

The OSCC gold standard dataset is likely to be of high quality and useful for future development of filters for studies of potential prognostic biomarkers. The methodology that we used is generalizable to other domains requiring a reference standard to compare performance of candidate filters. The labels regarding

relevancy in the gold standard dataset enable computation of diagnostic metrics, such as sensitivity and specificity. Thus, empirically derived measures of filter performance can guide iterative refinement. Librarians and informationists with an understanding of data analysis can contribute to the development of gold standard datasets and subsequent filters tuned for their patrons' domains of interest.

ACKNOWLEDGMENTS

We thank Michele Morris, systems developer, Department of Biomedical Informatics, University of Pittsburgh, for help in implementing GIANT, and Andrea Ketchum, AHIP, librarian, University of Pittsburgh Health Sciences Library System, for help in searching databases for reports of biomarker filter development.

REFERENCES

1. National Center for Biotechnology Information. PubMed.gov [Internet]. US National Library of Medicine, National Institutes of Health [cited 17 Feb 2014]. <http://www.ncbi.nlm.nih.gov/pubmed/>.
2. US National Library of Medicine. Fact sheet: MEDLINE [Internet]. Bethesda, MD: US National Institutes of Health [rev. 7 May 2014; cited 4 Jun 2014]. <http://www.nlm.nih.gov/pubs/factsheets/medline.html>.
3. Boissier MC. Benchmarking biomedical publications worldwide. Rheumatology (Oxford). 2013 Sep;52(9):1545–6.
4. Wilczynski NL, Haynes RB. Developing optimal search strategies for detecting clinically sound prognostic studies in MEDLINE: an analytic survey. BMC Med. 2004 Jun 9;2:23.
5. Haynes RB, McKibbon KA, Wilczynski NL, Walter SD, Werre SR. Optimal search strategies for retrieving scientifically strong studies of treatment from Medline: analytical survey. BMJ. 2005 May 21;330(7501):1179.
6. Jenkins M. Evaluation of methodological search filters—a review. Health Info Lib J. 2004 Sep;21(3):148–63.
7. Sampson M, Zhang L, Morrison A, Barrowman NJ, Clifford TJ, Platt RW, Klassen TP, Moher D. An alternative to the hand searching gold standard: validating methodological search filters using relative recall. BMC Med Res Methodol. 2006;6:33.
8. Wilczynski NL, Morgan D, Haynes RB. An overview of the design and methods for retrieving high-quality studies for clinical care. BMC Med Inform Decis Mak. 2005;5:20.
9. Mathet Y, Widlocher A, Fort K, Francois C, Galibert O, Grouin C, Kahn J, Rosset S, Zweigenbaum P. Manual corpus annotation: giving meaning to the evaluation metrics. International Conference on Computational Lingustics; 2012.
10. Pustejovsky J, Stubbs A. Natural language annotation for machine learning. Sebastopol, CA: O'Reilly Media; 2012.
11. Bird S, Klein E, Loper E. Natural language processing with Python. Sebastopol, CA: O'Reilly; 2009.
12. Neville BW, Damm DD, Allen CE, Bouquot JE. Oral and maxillofacial pathology. 3rd ed. St. Louis, MO: Saunders/Elsevier; 2009.
13. Shaw AT, Yeap BY, Solomon BJ, Riely GJ, Gainor J, Engelman JA, Shapiro GI, Costa DB, Ou SH, Butaney M, Salgia R, Maki RG, Varella-Garcia M, Doebele RC, Bang YJ, Kulig K, Selaru P, Tang Y, Wilner KD, Kwak EL, Clark JW, Iafrate AJ, Camidge DR. Effect of crizotinib on overall survival in patients with advanced non-small-cell lung cancer harbouring ALK gene rearrangement: a retrospective analysis. Lancet Oncol. 2011 Oct;12(11):1004–12.
14. Meric-Bernstam F, Farhangfar C, Mendelsohn J, Mills GB. Building a personalized medicine infrastructure at a major cancer center. J Clin Oncol. 2013 May 20;31(15):1849–57.
15. McShane LM, Hayes DF. Publication of tumor marker research results: the necessity for complete and transparent reporting. J Clin Oncol. 2012 Dec 1;30(34):4223–32.
16. Pratheepa L, Pratibha R, Sherlin HJ, Anuja N, Premkumar P. Expression of emerging novel tumor markers in oral squamous cell carcinoma and their clinical and pathological correlation to determine the prognosis and usefulness as a therapeutic target: a systematic review. J Natural Sciences Res. 2012;2(1)
17. Choong MK, Tsafnat G. The implications of biomarker evidence for systematic reviews. BMC Med Res Methodol. 2012;12:176.
18. Elsevier. EMBASE [Internet]. Elsevier [cited 8 Jun 2014]. <http://www.elsevier.com/online-tools/embase>.
19. EBSCO. Library, Information Science & Technology Abstracts (LISTA) [Internet]. EBSCO; 2014 [cited 8 Jun 2014]. <http://www.ebscohost.com/academic/library-information-science-and-technology-abstracts>.
20. InterTASC Information Specialists' Sub-Group (ISSG). ISSG search filters resource: the InterTASC Information Specialists' Sub-Group search filter resource [Internet]. The Sub-Group [cited 22 May 2014]. <http://sites.google.com/a/york.ac.uk/issg-search-filters-resource/home>.
21. US National Library of Medicine. Medical Subject Headings: MeSH browser [Internet]. Bethesda, MD: US National Institutes of Health; 2014 [cited 14 Feb 2014]. <http://www.nlm.nih.gov/mesh/2014/mesh_browser/MBrowser.html>.
22. van de Glind EM, van Munster BC, Spijker R, Scholten RJ, Hooft L. Search filters to identify geriatric medicine in Medline. J Am Med Inform Assoc. 2012 May–Jun;19(3):468–72.
23. Geersing GJ, Bouwmeester W, Zuithoff P, Spijker R, Leeflang M, Moons KG. Search filters for finding prognostic and diagnostic prediction studies in Medline to enhance systematic reviews. PLOS One. 2012;7(2):e32844.
24. Yao X, Wilczynski NL, Walter SD, Haynes RB. Sample size determination for bibliographic retrieval studies. BMC Med Inform Decis Mak. 2008;8:43.
25. EndNote v.X6 [computer program]. Thomson Reuters.
26. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977 Mar;33(1):159–74.
27. Wilczynski NL, McKibbon KA, Walter SD, Garg AX, Haynes RB. MEDLINE clinical queries are robust when searching in recent publishing years. J Am Med Inform Assoc. 2013 Mar 1;20(2):363–8.
28. Wilczynski NL, McKibbon KA, Haynes RB. Response to Glanville et al.: how to identify randomized controlled trials in MEDLINE: ten years on [letter to the editor]. J Med Lib Assoc. 2007 Apr;95(2):117–8. DOI: http://dx.doi.org/10.3163/1536-5050.95.2.117.
29. Glanville J, Lefebvre C. Authors' response [letter to the editor]. J Med Lib Assoc. 2007 Apr;95(2):119–20.
30. Funk ME, Reid CA. Indexing consistency in MEDLINE. Bull Med Lib Assoc. 1983 Apr;71(2):176–83.
31. Bekhuis T, Demner-Fushman D, Crowley RS. Comparative effectiveness research designs: an analysis of terms and coverage in Medical Subject Headings (MeSH) and Emtree. J Med Lib Assoc. 2013 Apr;101(2):92–100. DOI: http://dx.doi.org/10.3163/1536-5050.101.2.004.
32. Medelyan O, Witten IH. Measuring inter-indexer consistency using a thesaurus. Proceedings of the 6th Association for Computing Machinery (ACM) and the IEEE Computer Society (IEEE-CS) Joint Conference on Digital Libraries. ACM. 2006:274–5.

AUTHORS' AFFILIATIONS

John J. Frazier, DMD, MSPH, jjf60@pitt.edu, Fellow of the American Academy of Oral and Maxillofacial Pathology, Diplomate of the American Board of Oral and Maxillofacial Pathology, and National Library of Medicine Fellow; **Corey D. Stein, MS,** cds51@pitt.edu, Researcher; **Eugene Tseytlin, MS,** tseytlin@pitt.edu, Systems Developer; **Tanja Bekhuis, PhD, MS, MLIS, AHIP** (Featured)**,** tcb24@pitt.edu, Assistant Professor, Department of Biomedical Informatics and Department of Dental Public Health; School of Medicine and School of Dental Medicine, University of Pittsburgh, 5607 Baum Boulevard, Suite 514, Pittsburgh, PA 15206-3701

Received March 2014; accepted August 2014